Rebuttal Supplement

Table 1: Task predictive performance (accuracy or AUC in %) across all tasks and baselines when intervening on a fixed fraction of the set of available concepts at test time (indicated on the left-hand side of the dataset names). Except for "IntCEM (Learnt Policy ψ)", all interventions are done using a random intervention policy. We write in bold the best performance results across baselines following a random intervention policy.

	Dataset	IntCEM (Random)	CEM	Joint CBM-Sigmoid	Joint CBM-Logit	Ind. CBM	Seq. CBM	IntCEM (Learnt Policy ψ)
25%	MNIST-Add	**93.78 ± 0.09**	91.66 ± 0.41	86.47 ± 1.11	79.86 ± 3.12	84.11 ± 1.09	88.88 ± 0.67	94.75 ± 0.28
	MNIST-Add-Incomp	**90.88 ± 0.25**	88.22 ± 0.92	84.68 ± 0.83	77.59 ± 1.57	86.06 ± 0.77	87.53 ± 0.74	91.69 ± 0.39
	CUB	**88.53 ± 0.23**	84.08 ± 0.36	83.18 ± 0.22	81.76 ± 0.90	79.96 ± 1.57	65.40 ± 5.38	94.10 ± 0.49
	CUB-Incomp	**78.34 ± 0.34**	77.83 ± 0.33	60.94 ± 3.66	62.97 ± 7.19	48.16 ± 0.62	42.78 ± 1.63	81.71 ± 0.23
	CelebA	**42.25 ± 0.36**	33.27 ± 0.94	25.16 ± 0.58	20.02 ± 1.27	25.63 ± 0.55	26.21 ± 0.32	47.63 ± 1.49
50%	MNIST-Add	**95.70 ± 0.11**	93.36 ± 0.51	88.66 ± 1.20	80.69 ± 3.30	85.71 ± 0.94	90.87 ± 0.83	96.71 ± 0.20
	MNIST-Add-Incomp	**92.07 ± 0.21**	89.13 ± 0.96	85.73 ± 1.12	77.96 ± 1.73	86.87 ± 0.74	88.58 ± 0.95	92.98 ± 0.31
	CUB	**96.14 ± 0.35**	89.66 ± 1.53	92.28 ± 0.74	86.37 ± 1.18	92.19 ± 0.77	74.24 ± 5.68	99.17 ± 0.14
	CUB-Incomp	**85.86 ± 0.29**	82.32 ± 0.58	66.12 ± 3.81	56.30 ± 6.71	54.35 ± 0.35	47.62 ± 1.85	90.69 ± 0.30
	CelebA	**52.01 ± 0.45**	38.82 ± 1.32	27.19 ± 1.17	21.20 ± 2.74	27.79 ± 1.07	29.11 ± 0.40	62.01 ± 1.21
75%	MNIST-Add	**97.51 ± 0.08**	95.01 ± 0.57	91.24 ± 1.05	81.31 ± 3.44	86.97 ± 0.96	92.58 ± 0.82	98.22 ± 0.09
	MNIST-Add-Incomp	**93.48 ± 0.25**	90.35 ± 0.96	88.09 ± 1.03	78.62 ± 1.90	88.15 ± 0.68	90.09 ± 0.91	94.18 ± 0.24
	CUB	**98.98 ± 0.14**	93.95 ± 1.95	95.73 ± 1.00	91.18 ± 1.61	96.97 ± 0.53	79.81 ± 4.96	99.82 ± 0.08
	CUB-Incomp	**91.74 ± 0.19**	86.93 ± 0.61	71.54 ± 3.93	57.39 ± 6.36	61.92 ± 0.52	53.00 ± 1.76	94.96 ± 0.26
	CelebA	**56.99 ± 0.37**	41.59 ± 1.39	28.04 ± 1.42	23.70 ± 3.82	28.63 ± 1.48	30.04 ± 0.48	66.48 ± 1.10
100%	MNIST-Add	**99.51 ± 0.04**	96.68 ± 0.68	94.84 ± 0.75	81.92 ± 3.65	88.43 ± 1.03	94.58 ± 0.81	99.51 ± 0.04
	MNIST-Add-Incomp	**94.99 ± 0.11**	91.53 ± 1.01	91.28 ± 0.81	79.21 ± 1.98	89.36 ± 0.69	91.54 ± 1.03	94.99 ± 0.11
	CUB	**99.90 ± 0.04**	96.75 ± 1.72	96.42 ± 1.24	95.03 ± 1.95	98.97 ± 0.35	82.81 ± 3.72	99.90 ± 0.04
	CUB-Incomp	**96.52 ± 0.19**	91.47 ± 0.67	76.72 ± 4.14	65.25 ± 6.88	69.54 ± 1.05	58.19 ± 1.85	96.52 ± 0.19
	CelebA	**70.02 ± 0.62**	48.10 ± 1.72	29.38 ± 1.80	30.75 ± 9.89	29.43 ± 2.02	31.04 ± 0.22	70.02 ± 0.62

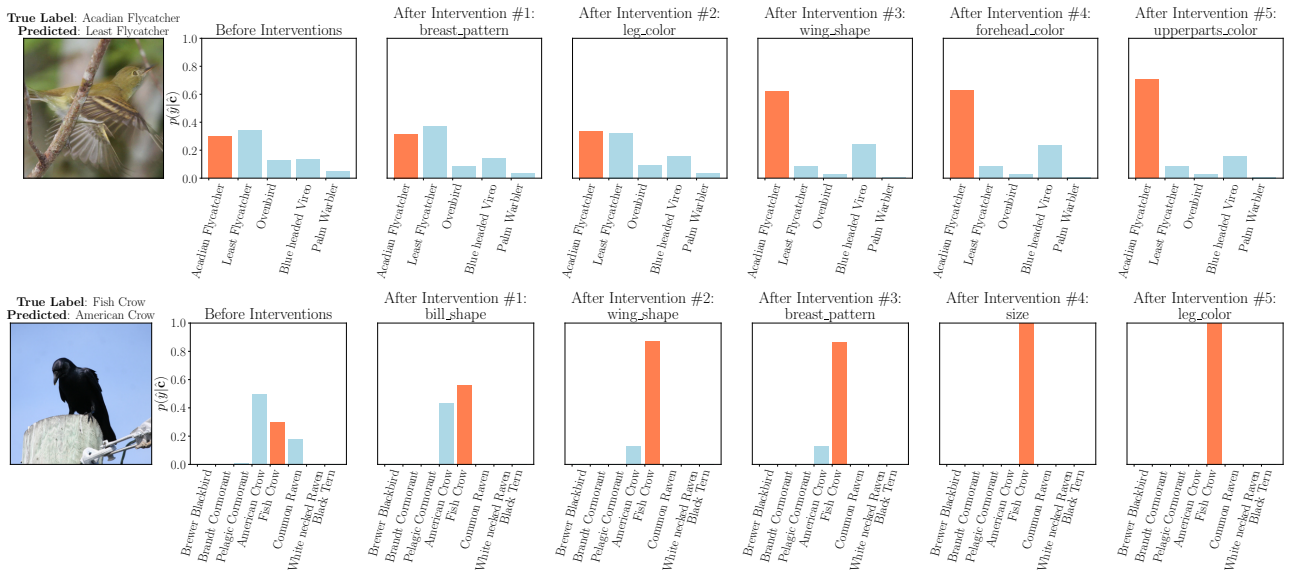

Figure 1: Examples of five consecutive concept interventions on an IntCEM following its learnt policy ψ for two random test samples in `CUB`. Each column indicates the posterior task label distribution $p(\hat{y}|\hat{\mathbf{c}})$ after each intervention, with the first panel showing the predictive task distribution before any concept interventions were made. We only show the classes with the highest probability and highlight the correct label's probability using orange bars. The concept selected by ψ for each sample at each intervention step is indicated above the panels.

Table 2: Oracle Impurity Score (OIS) for all jointly-trained baselines across all tasks. Higher OIS values indicate higher leakage in a model's learnt concept representations. Our results suggest that IntCEMs tend to have higher amounts of leakage than other models across all datasets (except for `CUB-Incomp`).

	Dataset	IntCEM	CEM	Joint CBM-Sigmoid	Joint CBM-Logit
OIS (%)	MNIST-Add	30.97 ± 0.29	28.25 ± 0.44	13.14 ± 0.27	20.96 ± 0.02
	MNIST-Add-Incomp	36.73 ± 0.23	34.09 ± 0.47	13.28 ± 0.27	26.12 ± 0.08
	CUB	45.86 ± 0.29	42.54 ± 2.30	21.01 ± 0.58	41.66 ± 0.49
	CUB-Incomp	38.79 ± 1.41	42.13 ± 3.48	27.81 ± 2.02	30.32 ± 3.09
	CelebA	50.87 ± 4.07	40.63 ± 4.83	29.65 ± 16.51	24.11 ± 09.30